Real-time Robust Automatic Speech Recognition Using Compact Support Vector Machines

Rubén Solera-Ureña*, *Member, IEEE,* Ana Isabel García-Moral, Carmen Peláez-Moreno, *Member, IEEE,* Manel Martínez-Ramón, *Senior Member, IEEE,* and Fernando Díaz-de-María, *Member, IEEE*

***Abstract*—In the last years, support vector machines (SVMs) have shown excellent performance in many applications, especially in the presence of noise. In particular, SVMs offer several advantages over artificial neural networks (ANNs) that have attracted the attention of the speech processing community. Nevertheless, their high computational requirements prevent them from being used in practice in automatic speech recognition (ASR), where ANNs have proven to be successful. The high complexity of SVMs in this context arises from the use of huge speech training databases with millions of samples and highly overlapped classes. This paper suggests the use of a weighted least squares (WLS) training procedure that facilitates the possibility of imposing a compact semiparametric model on the SVM, which results in a dramatic complexity reduction. Such a complexity reduction with respect to conventional SVMs, which is between two and three orders of magnitude, allows the proposed hybrid WLS-SVC/HMM system to perform real-time speech decoding on a connected-digit recognition task (SpeechDat Spanish database). The experimental evaluation of the proposed system shows encouraging performance levels in clean and noisy conditions, although further improvements are required to reach the maturity level of current context-dependent HMM-based recognizers.**

***Index Terms*—Robust ASR, additive noise, hybrid ASR, hidden Markov models (HMMs), machine learning, artificial neural networks (ANNs), support vector machines (SVMs), ANN/HMM, SVM/HMM, real-time ASR, compact SVM.**

I. Introduction

HIDDEN Markov models (HMMs) have become the most employed core technique for automatic speech recognition (ASR). However, the HMM-based ASR systems seem to be close to reaching their limit of performance. Hybrid systems based on a combination of artificial neural networks (ANNs) and HMMs, referred to as hybrid ANN/HMM [1]–[3], provide significant performance improvements in noisy conditions [4], [5]. However, progress on this paradigm has been hindered by their training computational requirements, which were excessive at the time these systems were proposed, and the inherent difficulty of competing with a technique that has been fine tuned during decades.

Support vector machines (SVMs) [6] have shown superior performance than ANNs in a variety of tasks. There are two fundamental reasons: first, the SVM training process is guaranteed to converge to the global minimum of the associated cost function; and second, SVMs exhibit superior generalization capability. This last property allows SVMs to make more accurate decisions in noisy environments, which is a valuable characteristic in the field of automatic speech recognition. Inspired by these potential strengths, several authors have suggested the use of SVMs in ASR [7]–[13]. However, a key difficulty still remains: though hybrid SVM/HMM systems [10]–[12] are able to deal with the time variability of speech utterances and reasonable solutions for multiclass classification and probability estimation have been proposed, the resulting SVMs are too complex and computationally demanding to allow for real-time speech recognition.

This problem is actually twofold: first, the maximum number of samples that can be used for the SVM training is limited to a few millions; second, large speech databases with highly overlapped classes lead to huge models that must be evaluated at the decoding phase. In this work, the first problem is alleviated by randomly selecting a balanced subset of training samples, which significantly reduces the computational cost of the training process while causing negligible reduction in performance, as it was previously demonstrated in the ANN/HMM paradigm [5]. However, further research on this issue is required for the proposed system to manage more demanding ASR tasks. It is the second of the above-mentioned problems that actually hinders the possibility of real-time decoding of speech utterances. This issue constitutes the focus of the present paper.

The complexity of SVMs in a hybrid SVM/HMM speech recognition system must be notably reduced in order to achieve a real-time operation. To this end, we propose here the use of compact SVMs. Specifically, we suggest training the SVMs through a weighted least squares (WLS) procedure [14] that converges to the original solution obtained by quadratic programming (QP) techniques. The WLS procedure does not produce any complexity reduction *per se*, but facilitates the possibility of selecting an *a priori* target complexity by imposing a compact semiparametric model on the SVM [15], [16], which is expressed in terms of a reduced set of representative vectors. To this end, a sequential selection approach based on the approximate linear dependence (ALD) condition [17], [18]

This work is partially supported by the projects funded by the Spanish Ministry of Science and Innovation TEC 2008-06382 and TEC 2008-02473 and by the regional grant Comunidad Autónoma de Madrid-UC3M CCG10-UC3M/TIC-5570.

R. Solera-Ureña, C. Peláez-Moreno, M. Martínez-Ramón and F. Díaz-de-María are with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Leganés - 28911, Spain (e-mail: rsolera@tsc.uc3m.es; carmen@tsc.uc3m.es; manel@tsc.uc3m.es; fdiaz@tsc.uc3m.es).

A. I. García-Moral is with Fonetic Solutions S. L., Madrid - 28037, Spain (e-mail: ana.garcia@fonetic.es).

*Corresponding author.

is employed to obtain a set of nearly independent vectors in the feature space. The use of these techniques leads to support vector classifiers (denoted as WLS-SVC) that are compact enough for the decoding system to operate in real-time on a medium-complexity speech recognition task, while maintaining its performance. Namely, the computational burden at the decoding stage, in terms of kernel evaluations, is reduced by between two and three orders of magnitude with respect to the baseline SVM/HMM system. Moreover, experimental results in clean and noisy conditions show similar or even better performance than standard monophone-based HMM ASR systems, while using only 13% of the full training dataset. Finally, it should be noted that the proposed WLS-SVC/HMM system constitutes a very promising starting point for the development of practical SVM-based ASR, but substantial improvements are still required. Specifically, both the training procedure when dealing with very large speech databases and the way to take full benefit of contextual information should be improved in order to reach the performance of state-of-the-art context-dependent HMM-based recognizers.

The rest of this paper is organized as follows. Hybrid systems for ASR are presented in Section II with special emphasis on the state-of-the-art of hybrid SVM/HMM speech recognition. Next, our proposal is described in Section III, which consists of a brief review of the WLS-SVC formulation and a description of the data selection methods employed to obtain a balanced subset of training samples and an adequate base of centroids for the compact SVM. Finally, experiments and results are presented in Section IV followed by conclusions and suggested future lines of research.

II. Hybrid Systems for Automatic Speech Recognition

A. Motivation

The discrimination ability of ANNs was soon recognized as a desirable characteristic that could contribute to the improvement of ASR systems. However, the duration variability of the speech instances corresponding to the same class hindered the straightforward application of ANNs. To overcome this problem, a variety of different architectures and novel training algorithms that combined HMMs with ANNs were proposed in the late 1980s and early 1990s. The fundamental advantage of this approach is the introduction of a discriminative technique (ANN) into a generative system (HMM) that retains its ability to handle the temporal variability of the speech signal. For a comprehensive survey of these techniques, see [3].

In this paper we have focused on the architecture, initially proposed by Bourlard and Morgan [1], [2], that applies ANNs to estimate the HMM emitting state likelihoods previously provided by Gaussian mixture models (GMMs). The authors exploited the well-known capability of feed-forward networks, such as multilayer perceptrons (MLPs), of estimating *a posteriori* probabilities when trained in classification mode (see [19] for the fundamentals of MLPs). The specific formulation will be introduced in Section II-B.

Though at the time this approach was suggested the use of these MLPs in speech recognition was still a challenging issue from a computational point of view, the following remarkable advantages were identified (from [20]):

- Model accuracy: ANNs have greater flexibility to provide more accurate acoustic models.
- Local discrimination ability (at a frame level): MLPs are trained to obtain class boundaries instead of providing an accurate (generative) model for each particular class.
- Parsimonious use of parameters: all the classes share the same ANN parameters (this does not hold for every ANN, but it does for MLPs).
- HMMs and ANNs exhibit complementary abilities for ASR tasks, which lead to higher recognition rates, especially under noisy conditions.
- Adaptation techniques have also been proposed (for example, speaker adaptation as in [21]–[23]).

Thanks to the improvement of computational capabilities, the last decade has witnessed an emergence of variants of this model that profited from the aforementioned advantages. In particular, hybrid systems have been found very appropriate and flexible for introducing all sorts of information missing in the classical HMM paradigm. From the parameterization point of view, features do not need to be uncorrelated because the network learns the local correlation between its input units. This has been used to include alternative features such as spectro-temporal parameters obtained by frequency filtering (FF) [4] or linear prediction [24], [25], or speech production knowledge in the form of articulatory features which led to more robust systems [26]–[28]. Most noteworthy, the possibility of augmenting the time-span in the feature extraction procedure together with the various methods available for combining these features (multistream, concatenation, probabilistic, etc.) has broadened the choices for phonetic context dependency representation [29], [30]. The addition of transitional units (diphones) is another economical alternative to triphone units for the inclusion of context-dependent acoustic modeling in the hybrid approach [31], [32].

As a drawback, we can mention that most implementations rely on an initial segmentation of the training set at the level of the classes considered by the ANN. That is, each training frame must have its corresponding class label (phoneme, state of phoneme, etc.). However, large databases are rarely manually labeled at a phoneme level because of the enormous human effort necessary for the task. Therefore, most state-of-the-art hybrid recognizers perform an initial forced alignment with conventional HMMs. This alignment becomes the ground truth for the training of the ANN. We have made use of this approach and further subdivided the phonemes into three sections (initial, middle, and final) making a finer segmentation attending to the distribution of the frames into the states of the HMMs employed for forced alignment. Further re-alignments using the models trained at each iteration would improve the segmentation of the training database, but this issue is beyond the scope of this work.

B. Problem Formulation

It is well known that the speech recognition problem can be stated as *finding the sequence of words $\tilde{W}$ that maximizes*

the probability $P(W|X)$, where $X = x_1, \ldots, x_T$ is the sequence of input observation features. This problem is usually factorized using Bayes' theorem as:

$$P(W|X) \propto P(X|W)\, P(W) \tag{1}$$

where the *a priori* probabilities $P(W)$ are modeled using a language model and the likelihoods $P(X|W)$ are estimated by the HMMs. Here, W is modeled as a sequence of states $W = q_1, \ldots, q_L$, where each state describes the probability of occurrence of an input feature vector $\mathbf{x}_t$ by means of an emission probability density function $p(\mathbf{x}_t|q_l)$. Hybrid SVM/HMM systems substitute Gaussian mixture models for support vector machines to provide robust *a posteriori* probabilities $p(q_l|\mathbf{x}_t)$, of class q_l given the feature vector $\mathbf{x}_t$. True emission likelihoods can be obtained from the probabilities provided by SVMs by using Bayes' rule:

$$\frac{p(\mathbf{x}_t|q_l)}{p(\mathbf{x}_t)} = \frac{p(q_l|\mathbf{x}_t)}{p(q_l)} \tag{2}$$

The *a priori* probability $p(\mathbf{x}_t)$ can be dropped from the equation as its value is the same for every class. Therefore, the *a posteriori* probabilities should be normalized by the class priors to obtain what are called *scaled likelihoods*. However, it will be noted later in this paper that such a normalization is unnecessary when balanced training datasets are used.

C. Hybrid SVM/HMM Systems

This section presents a description of the hybrid SVM/HMM systems proposed in the last years and their practical limitations, which justify, in our opinion, the interest of the work presented in this paper. For a more detailed review on the use of support vector machines for ASR, including systems and difficulties, refer to [33], [34].

Hybrid systems based on discriminative models like artificial neural networks have demonstrated good performance in automatic speech recognition. Nevertheless, support vector machines offer several theoretical advantages that have attracted the attention of many speech processing researchers in the last years. Firstly, they are capable of dealing with samples of a very high dimensionality. Secondly, their convergence to the global minimum of the cost function is guaranteed by means of QP techniques. Finally, the maximum margin solution provides SVMs with superior generalization capability, which should result in improved robustness in the presence of noise. In our opinion, these characteristics make SVMs a promising future alternative to Gaussian mixture models and artificial neural networks for the problem of acoustic modeling in robust speech recognition.

However, the application of support vector machines to automatic speech recognition is not straightforward. There are mainly two reasons for the scarce use of SVMs in this field. First, the high computational cost of SVMs and their difficulty to handle large databases prevent them from being used in speech recognition. Second, SVMs are static classifiers that need fixed-dimension input vectors, so they cannot directly deal with the variable time duration of speech units.

The first problem has been avoided or even ignored in the great majority of works in the field, whereas several solutions can be found in the literature for the latter. Some of them perform a previous processing of the speech or feature sequence in order to obtain fixed dimension vectors that fit the SVM input. This normalization can be achieved by means of simple uniform [35] or non-uniform [36], [37] feature sequence resampling procedures. Other authors apply the so-called *triphone model approach*, which assumes that speech segments corresponding to phones are composed of a fixed number of sections (3 in most cases). Feature vectors in each segment are averaged and the results are then concatenated to form a fixed-dimension vector [38]–[40].

Preliminary versions of the hybrid SVM/HMM systems currently employed were proposed in [9], [41], [42] and [43], all of them comprising a two step decoding process. In [9], [41], [42], hidden Markov models are used to generate phonetic level alignments on the speech utterance. The previously mentioned triphone model approach is then applied to extract fixed-dimension feature vectors from each segment. On the other hand, the system described in [43] operates on a frame by frame basis. In both cases, the support vector classifier uses the feature vectors obtained in the previous step to generate segmental or instantaneous phoneme decisions, which are incorporated into a Viterbi decoding stage that rescores a N-best list provided by Gaussian mixture models in the first step.

A major drawback of all of these systems is the requisite for a previous segmentation of the speech utterances. In contrast to current one-pass hybrid systems, these speech recognizers require an HMM-based forced alignment in both training and recognition phases to achieve such a segmentation. This fact makes the practical application of support vector machines in automatic speech recognition difficult, since a double decoding process must be done. To overcome this problem, some authors proposed the combination of SVMs and HMMs in hybrid systems inspired by the ANN/HMM framework [1], [2]. The basis of this approach is to merge SVMs and HMMs into a single hybrid SVM/HMM system that benefits from their complementary abilities for ASR tasks, namely: the capability of HMMs to handle the time variability of speech and the discrimination power provided by support vector machines.

Hybrid SVM/HMM systems like those proposed in [10], [11], [44] replace Gaussian mixture models with support vector machines as probabilistic estimators in the acoustic modeling phase. Thus, SVMs estimate the HMM state emission probabilities that will be employed by a Viterbi decoder to obtain the transcription of the speech utterance. These systems work on a frame by frame basis and therefore do not need a previous segmentation of the speech utterances in the decoding stage, which is performed in a single step. In this case, only an initial state level alignment of the training set is required to obtain labeled feature vector examples for the multiclass SVM to be trained. A conventional GMM/HMM-based system is used for that purpose.

These systems achieve a similar or even slightly better performance than standard baseline HMM-based systems in clean conditions. Nonetheless, the high computational cost of SVMs has prevented them from becoming a viable alternative to conventional systems for robust automatic speech recognition.

The computational burden of support vector machines affects hybrid SVM/HMM systems in two ways. Firstly, it limits to a few millions the maximum number of samples that can be used in the training stage of the SVM. Secondly, large speech databases with highly overlapped classes lead to huge models, with too many support vectors whose kernel functions must be evaluated at the decoding phase. The former problem could limit the performance of the hybrid SVM/HMM recognition system, whereas the latter hinders a real-time decoding of the speech utterances.

Up to our knowledge, only the work in [12], [45] has addressed in a systematic manner the issue of the computational burden involved in hybrid SVM/HMM systems. This work focuses on identifying the indispensable binary classifiers, among those that form the multiclass SVM, that should be evaluated during the decoding stage to obtain accurate enough acoustic decisions. Thus, a dynamic selection method that picks out the most relevant binary SVMs and discards those less influential for the decision is proposed. On average, this method enables the evaluation of only 14% of the binary classifiers and reduces the recognition time between 90 and 180 times, with no performance degradation. However, this system still operates five-to-ten times slower than real-time speech recognition.

III. Real-time Hybrid SVM/HMM Automatic Speech Recognition

This section is devoted to the presentation of a new hybrid system that constitutes a promising starting point for the development of real-time SVM-based robust automatic speech recognition. This work is based on that described in [11], where a preliminary hybrid SVM/HMM system takes advantage of the discrimination power provided by SVMs to estimate robust emission probabilities, while keeping the capability of HMMs to handle the variable time duration of speech utterances. The drawback of its high computational burden at the decoding stage is faced in this work, with the result of a speech recognition system that is now capable of performing real-time speech decoding on a medium-complexity ASR task while achieving similar or even better results than context-independent HMM-based recognizers.

The approach proposed in this paper is based on the use of a compact semiparametric model for the SVM. A WLS procedure is then used to train the compact SVM. This procedure is carefully described in this section, in addition to some other practical issues related to the implementation of the hybrid SVM/HMM system.

A. Support Vector Machines

The support vector machine is a well-known statistical learning method, first proposed in [46] as an extension of the generalized portrait method for the construction of non-linear classifiers and regressors. Its formulation is based on Statistical Learning Theory (SLT) and implements the structural risk minimization (SRM) criterion [47], a principle that bounds overfitting by setting a trade-off between the model complexity and its empirical risk. This leads to the *maximum margin* solution, which endows the SVM with a higher generalization ability and, presumably, improved robustness in the presence of noise compared to other machine learning methods.

The support vector classifier (SVC) assigns a label $y \in \{\pm 1\}$ to the input vector $\mathbf{x}$ according to the following function:

$$f(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + b \tag{3}$$

where $\boldsymbol{\phi}(\mathbf{x}) : \Re^d \mapsto \Re^H$ is a nonlinear function that maps input vector $\mathbf{x}$ into a feature space of a higher (possibly infinite) dimensionality. The vector $\mathbf{w}$ denotes the separating hyper-plane in such a space and b represents the bias with respect to the origin.

The reason that gives the SVM good generalization properties is that its formulation involves a joint minimization of both empirical and structural risks. Structural risk minimization is equivalent to the minimization of the norm of vector $\mathbf{w}$. Thus, the solution to the SVM is given by the minimization of the following quadratic problem:

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i, \\ \text{subject to} \quad & y_i\left(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + b\right) \geq 1 - \xi_i; \;\; \forall i = 1,\ldots,n \\ & \xi_i \geq 0; \;\; \forall i = 1,\ldots,n \end{aligned} \tag{4}$$

where $\mathbf{x}_i \in \Re^d \;\; (i = 1,\ldots,n)$ are the training vectors with labels $y_i \in \{\pm 1\}$. The variables ξ_i represent the error for every input vector and C sets the compromise between the minimization of empirical and structural risks.

This problem is usually solved using the Wolfe dual [48], where Lagrange multipliers α_i are found according to:

$$\begin{aligned} \max_{\alpha_i} \quad & \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j \boldsymbol{\phi}^T(\mathbf{x}_i)\,\boldsymbol{\phi}(\mathbf{x}_j) \\ \text{subject to} \quad & \sum_{i=1}^{n}\alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C; \;\; \forall i = 1,\ldots,n \end{aligned} \tag{5}$$

The optimum decision boundary $\mathbf{w}$ is given by:

$$\mathbf{w} = \sum_{i=1}^{n}\alpha_i y_i \boldsymbol{\phi}(\mathbf{x}_i) \tag{6}$$

Only those training vectors with associated Lagrange multiplier $\alpha_i \neq 0$ will contribute to determining the decision boundary, thus receiving the name of *support vectors*. The mapping function $\boldsymbol{\phi}(\mathbf{x})$ is seldom explicitly known. However, the optimization problem in (5) is set in terms of dot products $\boldsymbol{\phi}^T(\mathbf{x}_i)\,\boldsymbol{\phi}(\mathbf{x}_j)$, which can be evaluated using a Mercer *kernel function* $K(\cdot,\cdot)$. The Mercer Theorem [49] states that a mapping function $\boldsymbol{\phi}$ and a function $K(\mathbf{x}_i,\mathbf{x}_j) = \boldsymbol{\phi}^T(\mathbf{x}_i)\,\boldsymbol{\phi}(\mathbf{x}_j)$ exist if and only if $K(\cdot,\cdot)$ is positive semidefinite. By means of the so-called *kernel trick*, the output of the SVM finally adopts the following expression:

$$f(\mathbf{x}) = \sum_{i=1}^{n}\alpha_i y_i K(\mathbf{x}_i,\mathbf{x}) + b \tag{7}$$

B. Multiclass SVMs and Probability Estimation

Support vector machines are binary classifiers in their original formulation, whereas the acoustic modeling stage in ASR can be stated as a multiclass problem. Nevertheless, there exist several ways to solve k-class problems using SVMs. *True* multiclass solutions reformulate the SVM equations to consider all classes at once in a single optimization problem [50]–[52]. Other methods are based on the combination of a number of binary classifiers, each of them trained independently from the others. On the one hand, the *one-versus-the-rest* approach trains k binary classifiers that compare each class against the rest. To test for a new vector all the classifiers are evaluated and the test sample is assigned to the classifier (class) with the largest output. On the other hand, the *one-versus-one* method trains $\frac{k(k-1)}{2}$ binary classifiers, each of them comparing two classes. In the test phase, all of the classifiers are evaluated and then a voting scheme or a multiclass probability estimation method is adopted to assign the test sample to its corresponding class. Besides these methods, there exist other multiclass approaches such as the directed acyclic graph [53] or the error-correcting output codes [54], [55], less used in practice.

The choice of a suitable multiclass SVM method heavily depends on the specific characteristics of the problem at hand. Both the size of the database and the complexity of the speech recognition task addressed in this work advise using the *one-versus-one* approach. Several arguments support this statement. Firstly, this method is preferred when dealing with large training datasets (see [56] for a detailed discussion) due to the fact that SVM's computational burden at the training step is approximately quadratic in the number of samples. The computational load of the task addressed in this paper is too large (in terms of memory requirements) for *one-versus-the-rest* and *true* multiclass SVMs, as they must handle some million training samples at once. In contrast, although the *one-versus-one* method must train more binary classifiers than the other approaches, each classifier is trained with a smaller fraction of the database. Secondly, each binary classifier in the *one-versus-one* approach deals with a more simple, balanced and easily separable problem. Finally, the reduction of the whole multiclass problem into smaller binary classification tasks allows for the use of larger training datasets, which provide more varied acoustical information for the speech recognition task.

In this work, the classes considered by the support vector machine correspond to the states of the phoneme hidden Markov models. As will be shown later in this paper, 18 Spanish context-independent phonemes are modeled by 3-state HMMs, which leads to 54 acoustic classes. Thus, 1431 binary classifiers must be trained according to the *one-versus-one* approach.

A multiclass support vector machine is used in the hybrid approach to estimate HMM-state emission probabilities. SVMs do not directly provide calibrated posterior probabilities but class labels. Nevertheless, several methods have been proposed to obtain these probabilities from SVM outputs. One of the most widely employed when dealing with multiclass problems, which is implemented by the LibSVM toolbox [57] used in this work, is based on the calculation of Platt's probabilities [58] for every binary classifier. This method assumes roughly exponential class-conditional densities between the margins in each binary classifier. Bayes' rule on two exponentials suggest using a sigmoidal parametric model for the posterior probability. Thus, assuming that a *one-versus-one* multiclass approach is used, Platt's probabilities of $\mathbf{x}$ belonging to class i are calculated for every binary SVM (i,j) as follows:

$$\begin{aligned} r_{ij}\left(\mathbf{x}\right) =& p\left(y=i|y=i \text{ or } j, f_{ij}\left(\mathbf{x}\right)\right) = \\ =& \frac{1}{1+\exp\left(a_{ij}f_{ij}\left(\mathbf{x}\right)+b_{ij}\right)} \\ r_{ji}\left(\mathbf{x}\right) =& p\left(y=j|y=i \text{ or } j, f_{ij}\left(\mathbf{x}\right)\right) = 1 - r_{ij}\left(\mathbf{x}\right) \end{aligned} \quad (8)$$

where $f_{ij}\left(\mathbf{x}\right)$ is the output of the binary classifier (i,j) for sample $\mathbf{x}$. The sigmoid's parameters a_{ij} and b_{ij} are estimated discriminatively by maximizing the log-likelihood function over training data.

These *binary* probabilities $r_{ij}\left(\mathbf{x}\right)\ \forall i,j=1,\ldots,k$ must be translated into *multiclass a posteriori* probabilities $p\left(q_i|\mathbf{x}\right) = p\left(y=i|\mathbf{x}\right)\ \forall i = 1,\ldots,k$. To this end, a version of the Refregier-Vallet method based on the Bradley-Terry model is used [59]. The following optimization problem must be solved once for each input pattern $\mathbf{x}$ in order to obtain the corresponding posterior probabilities:

$$\begin{aligned} \min_{\mathbf{p}} \ & \frac{1}{2}\sum_{i=1}^{k}\sum_{j:j\neq i}^{k}\left(r_{ji}p\left(q_i|\mathbf{x}\right) - r_{ij}p\left(q_j|\mathbf{x}\right)\right)^2 \\ \text{subject to } & \sum_{i=1}^{k} p\left(q_i|\mathbf{x}\right) = 1 \\ & p\left(q_i|\mathbf{x}\right) \geq 0; \quad \forall i = 1,\ldots,k \end{aligned} \quad (9)$$

where $\mathbf{p}\left(\mathbf{x}\right) = \left[p\left(q_1|\mathbf{x}\right),\ldots,p\left(q_k|\mathbf{x}\right)\right]^T$. This problem is convex and can be solved by means of a simple iterative method.

C. Data Selection and Balancing

Software tools employed at present to train support vector machines can only deal with a maximum of a few million training samples. In the state-of-the-art HMM-based framework, however, large databases containing several hundred hours of recorded speech have become an indispensable basis for relevant performance improvements. This makes the research in the hybrid SVM/HMM framework extremely difficult due to the huge computer memory requirements of SVMs and the large amount of time spent on tuning, training and testing these models. Thus, a reduction of the size of the datasets employed for training the SVMs becomes essential.

It is worth mentioning that such a reduction should be done by taking into account the particular characteristics of the speech database. Namely, the non-uniform distribution of the sounds of a given language and their different temporal durations lead to highly imbalanced classes. This means that certain phonemes are overrepresented in the speech databases in comparison to others, which results in skewed classification problems. In our case, two main consequences of imbalanced

data can be stressed. First, especial care must be taken with respect to the minority classes, as scarce and/or short phonemes are usually the key to distinguish among confusable sets of words. Second, highly imbalanced problems can bias the solution obtained by the support vector machine to the most populated class.

Data selection is a common practice among the machine learning community, where several techniques have been proposed in the last years to deal with imbalanced data (see [60]–[62] for an overview). However, the need for whole speech training utterances complicates the application of these techniques in conventional HMM-based ASR. This problem is overcome by hybrid systems, where theoretically i.i.d. training samples are presented individually to the classifier. Some practical examples in the ANN/HMM context can be found in [63], [64].

In this work a simple selection method based on a random downsampling of the majority classes in the whole original training database is used to produce fully-balanced reduced training sets. As a result, all classes (states of phonemes) are represented by the same number of training samples, which is given by the less populated class. Despite being a straightforward solution, the balanced approach presents several beneficial consequences. First, and most important, it reduces the computational burden in the training stage significantly without a loss of performance. Second, it overcomes the problem of training the support vector machine with imbalanced datasets that may affect the determination of the optimal decision boundary. Finally, this simple solution provides the desired emission likelihoods as the outputs of the SVM. The problem of obtaining *scaled likelihoods* from *a posteriori* probabilities in the hybrid ANN/HMM context was an open issue since mismatches between the *a priori* probabilities of the training and test databases led to inconsistent results [1], [4], [20], [63], [65], [66]. In [5] it is shown that *scaled likelihoods* should always be estimated using the prior probabilities from the training data. In our case, the balancing of the training set enables the interpretation of the outputs of the SVM as *scaled likelihoods* without the need of applying any corrections.

D. Review of the WLS-SVC Formulation

The computational burden of SVMs at the decoding stage depends on the number of support vectors, that is, those training samples that are present in (7) with $\alpha_i \neq 0$. In the standard SVM formulation, support vectors are given by the resolution method. This leads to huge machines that cannot be used in real complex applications such as speech recognition, where one can find a large number of training samples distributed through a number of highly overlapped classes.

This drawback can be overcome by means of an alternative training procedure [14], which solves a series of weighted least squares problems that converges to the support vector classifier solution. This method, called WLS-SVC, does not produce any complexity reduction *per se*. However, it is more versatile than traditional QP schemes and, additionally, facilitates the possibility of developing compact solutions through the use of a preset compact model for the SVM [15], [16]. A brief description of the WLS-SVC algorithm is presented below. A more detailed derivation of the mathematical formulation can be found in [14], [15], [67]. Furthermore, its convergence to the original SVC solution is proven in [68].

Let us revisit the primal formulation defined by (4). The linear constraints in that expression can be incorporated into the so-called Lagrangian functional with associated Lagrange multipliers α_i and μ_i, respectively:

$$L_P = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n}\xi_i\left(C-\alpha_i-\mu_i\right) + \\ + \sum_{i=1}^{n}\alpha_i\left[1-y_i\left(\mathbf{w}^T\boldsymbol{\phi}\left(\mathbf{x}_i\right)+b\right)\right] \tag{10}$$

The second term in (10) vanishes as KKT conditions must hold (see references [69] and [70], page 131, for more details). After several operations, the Lagrangian can be seen as a weighted least squares functional plus a Tikhonov regularization term [71]:

$$L_P = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{2}\sum_{i=1}^{n}a_i {e_i}^2 \tag{11}$$

where:

$$a_i = \frac{2\alpha_i}{1-y_i\left(\mathbf{w}^T\boldsymbol{\phi}\left(\mathbf{x}_i\right)+b\right)} = \frac{2\alpha_i}{e_i y_i} \tag{12}$$

and $e_i = y_i - \left(\mathbf{w}^T\boldsymbol{\phi}\left(\mathbf{x}_i\right)+b\right)$ is the error for the training vector $\mathbf{x}_i$.

The minimization of (11) with respect to $\mathbf{w}$ and b cannot be done in a single step because a_i depends on $\mathbf{w}$. Thus, the following iterative WLS procedure was proposed in [14]:

1) Minimize (11) with respect to $\mathbf{w}$ and b, assuming that a_i holds fixed.
2) Update a_i using e_i and KKT conditions.
3) Repeat until convergence.

The minimization of (11) produces the following system:

$$\begin{bmatrix}\boldsymbol{\Phi}\mathbf{D_a}\boldsymbol{\Phi}^T+\mathbf{I} & \boldsymbol{\Phi}\mathbf{a}\\ \mathbf{a}^T\boldsymbol{\Phi}^T & \mathbf{a}^T\mathbf{1}\end{bmatrix}\begin{bmatrix}\mathbf{w}\\ b\end{bmatrix} = \begin{bmatrix}\boldsymbol{\Phi}\mathbf{D_a}\mathbf{y}\\ \mathbf{a}^T\mathbf{y}\end{bmatrix} \tag{13}$$

where $\boldsymbol{\Phi} = [\boldsymbol{\phi}\left(\mathbf{x}_1\right)|\ldots|\boldsymbol{\phi}\left(\mathbf{x}_n\right)]$, $\mathbf{a} = [a_1,\ldots,a_n]^T$ and $\mathbf{D_a}$ is a diagonal matrix with $\left(\mathbf{D_a}\right)_{ii} = a_i\ \forall i = 1,\ldots,n$.

The solution $\begin{bmatrix}\mathbf{w}^T & b\end{bmatrix}^T$ of the above system of equations is expressed in terms of the nonlinear mapping function $\boldsymbol{\phi}\left(\mathbf{x}\right)$, which is seldom explicitly known. Fortunately, the representer theorem [72] states that vector $\mathbf{w}$ can be expressed as a linear combination of the training samples:

$$\mathbf{w} = \sum_{i=1}^{n}\beta_i\boldsymbol{\phi}\left(\mathbf{x}_i\right) = \boldsymbol{\Phi}\boldsymbol{\beta} \tag{14}$$

where $\boldsymbol{\beta} = [\beta_1,\ldots,\beta_n]^T$. Replacing its expression in (13) and after several algebraic transformations (see [67], Appendix A.1) follows:

$$\begin{bmatrix}\mathbf{K}+{\mathbf{D_a}}^{-1} & \mathbf{1}\\ \mathbf{1}^T & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}\\ b\end{bmatrix} = \begin{bmatrix}\mathbf{y}\\ 0\end{bmatrix} \tag{15}$$

where $\mathbf{K} = \mathbf{\Phi}^T\mathbf{\Phi}$ is the kernel matrix. The solution $\left[\boldsymbol{\beta}^T \; b\right]^T$ of the above system of equations must be obtained in every step of the WLS procedure. The value of the coefficients a_i can be obtained by forcing KKT conditions to hold. Knowing that:

$$\xi_i = \begin{cases} 0 & \text{if } e_i y_i < 0 \\ e_i y_i & \text{if } e_i y_i \geq 0 \end{cases} \tag{16}$$

a_i can be obtained as follows:

$$a_i = \begin{cases} 0 & \text{if } e_i y_i < 0 \\ \frac{2C}{e_i y_i} & \text{if } e_i y_i \geq 0 \end{cases} \tag{17}$$

In practice, a maximum value for a_i is imposed to avoid numerical problems when $e_i y_i$ goes to zero. This limitation is equivalent to a numerical regularization of the kernel matrix.

The output of the WLS-SVC described above adopts the following expression:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \beta_i K(\mathbf{x}_i, \mathbf{x}) + b \tag{18}$$

where it is worth noting that β_i converges asymptotically to $\alpha_i y_i$ in (7) and, therefore, the WLS-SVC converges to the original SVM.

Unlike with standard QP training methods, we can take advantage of the WLS formulation to fix *a priori* the complexity of the support vector machine. Compact machines can be implemented by imposing an alternative simple semiparametric model on vector $\mathbf{w}$:

$$\mathbf{w} = \sum_{i=1}^{r} \gamma_i \boldsymbol{\phi}(\mathbf{c}_i) = \mathbf{\Psi}\boldsymbol{\gamma} \tag{19}$$

where $\mathbf{\Psi} = [\boldsymbol{\phi}(\mathbf{c}_1) | \ldots | \boldsymbol{\phi}(\mathbf{c}_r)]$, $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_r]^T$ and $r << n$. Vectors $\mathbf{c}_i$ should form an orthogonal base for the training samples in the feature space. As the calculation of such a base can be hard, iterative sample selection methods, clustering techniques or PCA analysis can be used to select a set of representative centroids for the training database. In this case, the compact WLS-SVC solution obtained is just an approximation of the original SVC:

$$f(\mathbf{x}) = \sum_{i=1}^{r} \gamma_i K(\mathbf{c}_i, \mathbf{x}) + b \tag{20}$$

It should be noted, however, that the complexity of the compact WLS-SVC is not given by the number of support vectors anymore, but by the number of centroids.

By substituting (19) in (13), multiplying it by $\begin{bmatrix} \mathbf{\Psi}^T & \mathbf{0} \\ \mathbf{0}^T & 1 \end{bmatrix}$ and reordering terms, the following system is obtained:

$$\begin{bmatrix} \mathbf{K}_{\mathbf{\Phi}}^T \mathbf{D}_{\mathbf{a}} \mathbf{K}_{\mathbf{\Phi}} + \mathbf{K}_{\mathbf{\Psi}} & \mathbf{K}_{\mathbf{\Phi}}^T \mathbf{a} \\ \mathbf{a}^T \mathbf{K}_{\mathbf{\Phi}} & \mathbf{a}^T \mathbf{1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\gamma} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{K}_{\mathbf{\Phi}}^T \mathbf{D}_{\mathbf{a}} \mathbf{y} \\ \mathbf{a}^T \mathbf{y} \end{bmatrix} \tag{21}$$

where $\mathbf{K}_{\mathbf{\Phi}} = \mathbf{\Phi}^T\mathbf{\Psi}$ is the kernel matrix of the training samples and the centroids and $\mathbf{K}_{\mathbf{\Psi}} = \mathbf{\Psi}^T\mathbf{\Psi}$ is the kernel matrix of the centroids. This system of equations must be solved with respect to the solution $\left[\boldsymbol{\gamma}^T \; b\right]^T$ in every step of the WLS procedure, assuming fixed values for a_i. Next, their values must be updated according to (17).

The experimental results in Section IV-C show that efficient base selection methods, as that described in detail in Section III-E, enable important complexity reductions when using the compact support vector machine (up to 500 times and even higher), without a significant decrease in its recognition accuracy. Thus, the practical interest of the compact hybrid WLS-SVC/HMM system is demonstrated, achieving real-time speech decoding in a connected-digit recognition task with similar performance to that of the baseline SVM/HMM recognition systems.

E. Base Selection for Compact Multiclass SVMs

The hybrid SVM/HMM speech recognition systems described in Section II-C are still far from performing a real-time decoding. The main reason is that hard problems like automatic speech recognition, with millions of training samples and highly overlapped classes, result in huge support vector machines when using conventional QP training techniques. To alleviate this drawback, we suggest controlling the complexities of the SVMs by imposing a semiparametric compact model on the weight vector $\mathbf{w}$, as shown in Eq. (19). A WLS procedure is then used to train the compact SVM.

The key point, therefore, lies in finding a reduced yet representative set of centroids for the compact WLS-SVC. As previously stated, there exist a number of alternative procedures to the exhaustive search for an orthogonal base of vectors, which may be a hard problem. Several techniques such as clustering or PCA have been employed in other application contexts to obtain suitable bases for the semiparametric model [16].

In this work, a sequential selection approach based on the approximate linear dependence (ALD) condition is used to obtain a set of nearly independent vectors in the feature space. This procedure has been designed to exploit the specific distribution of the training samples in the feature space and the *one-versus-one* multiclass architecture employed in this work. Firstly, our selection method aims at reducing both intra-class redundancy and inter-class overlap in order to achieve a small base of representative centroids. The origin of this overlap is twofold: the coarticulation effects that make the boundaries of the classes quite blurry, which is augmented by the fact of using three different classes per phoneme, and the segmentation errors produced by the baseline HMM-based system that performs the initial forced alignment. Secondly, remarkable complexity reductions are achieved by forcing all the binary classifiers in the multiclass SVM to share a unique small base of centroids. It may seem surprising that the set of centroids of a given binary classifier can contain samples belonging to other different classes. This fact is explained, however, by the overlap existing between different classes that allows certain training samples to be represented by feature vectors from other classes.

It is worth highlighting the use of a unique base of centroids for the multiclass SVM, as it allows us to achieve larger complexity reductions than other methods previously published in the literature. For example, a similar approach based

on approximating the SVM decision surface by a reduced set expansion was proposed by Burges in [73]. Although this method has demonstrated good performance for binary classification tasks in the speech recognition framework [74], it fails to achieve the large complexity reductions required in this work since different support sets are obtained for each binary classifier. The same conclusion holds for the kernel online algorithm by Orabona et al. [75], where each support vector is shared by only two binary classifiers.

Basing on the ALD condition, the training samples are sequentially added to the base of centroids if their projection error exceeds a preset threshold. Specifically, given a set of centroids $\{\mathbf{c}_1, \ldots, \mathbf{c}_m\}$ and a new training sample $\mathbf{x}$, there exists an optimal linear combination of the elements of the base, with projection coefficients o_i, that minimizes the following squared error:

$$\delta = \min_{\mathbf{o}} \left\| \sum_{i=1}^{m} o_i \boldsymbol{\phi}(\mathbf{c}_i) - \boldsymbol{\phi}(\mathbf{x}) \right\|^2 \tag{22}$$

Solving (22) yields the optimal value of $\mathbf{o}$ [18]:

$$\mathbf{o} = \mathbf{K}_{\boldsymbol{\Psi}}^{-1} \mathbf{k}_{\boldsymbol{\Phi}} \tag{23}$$

and the residual:

$$\delta = K(\mathbf{x}, \mathbf{x}) - \mathbf{k}_{\boldsymbol{\Phi}}^T \mathbf{o} \tag{24}$$

where $\mathbf{K}_{\boldsymbol{\Psi}}$ is the kernel matrix of the m centroids in the base and $\mathbf{k}_{\boldsymbol{\Phi}}$ denotes the kernel vector of the centroids and the training sample $\mathbf{x}$: $(\mathbf{k}_{\boldsymbol{\Phi}})_i = K(\mathbf{c}_i, \mathbf{x})\ \forall i = 1, \ldots, m$. In our selection approach, a new training sample $\mathbf{x}$ will be added to the base of centroids if $\delta > \nu_g$, where ν_g is a preset accuracy threshold (*growing threshold*). Otherwise, the sample will not be added to the base of centroids as it can be represented with a negligible approximation error by the current m centroids.

Once the key fundamentals of the base selection method used in this work have been presented, we proceed to describe it in more detail. The selection procedure consists of the following sequential sample addition (growing) and deletion (pruning) processes:

Intra-class selection. Firstly, samples belonging to each class are processed independently to obtain a set of centroids $\mathbf{C}_{class}^{(i)}$ for every class $i = 1, \ldots, k$. This process aims at reducing the intra-class redundancy and consists of the following steps:

- A temporary base $\mathbf{C}_{class\_t}^{(i)}$ is initialized with the first training sample in class i.
- Training samples belonging to class i are processed sequentially in order to compute their ALD residual δ (24) with respect to $\mathbf{C}_{class\_t}^{(i)}$. A new training sample will be added to the temporary base if its residual δ is greater than the *growing threshold* ν_g.
- The projection coefficients vector $\mathbf{o}$ is computed for every training sample in class i. Their absolute values are then accumulated in a variable denoted by $\mathbf{o}_{acc1}$.
- The components of $\mathbf{o}_{acc1}$ are normalized with respect to the maximum. Those centroids in $\mathbf{C}_{class\_t}^{(i)}$ with an accumulated projection coefficient $(o_{acc1})_i$ lower than a *pruning threshold* ν_p are removed from the list. The remaining vectors will form the base $\mathbf{C}_{class}^{(i)}$ for class i.

Inter-class selection. Secondly, the centroids in all the k bases $\mathbf{C}_{class}^{(i)}$ are put together in a single temporary base $\mathbf{C}_{total}$ and then processed to eliminate the inter-class overlap. This procedure is similar to the previous one and consists of the following steps:

- A temporary base $\mathbf{C}_{total\_t}$ is initialized with the first vector in $\mathbf{C}_{total}$.
- The centroids in $\mathbf{C}_{total}$ are processed sequentially in order to compute their ALD residual δ (24) with respect to $\mathbf{C}_{total\_t}$. A new centroid will be added to $\mathbf{C}_{total\_t}$ if its residual δ is greater than the *growing threshold* ν_g.
- The projection coefficients vector $\mathbf{o}$ is computed for every sample in $\mathbf{C}_{total}$. Their absolute values are then accumulated in a variable denoted by $\mathbf{o}_{acc2}$.
- The components of $\mathbf{o}_{acc2}$ are normalized with respect to the maximum. Those centroids in $\mathbf{C}_{total\_t}$ with an accumulated projection coefficient $(o_{acc2})_i$ lower than the *pruning threshold* ν_p are removed from the list. The remaining vectors will form the definitive base of centroids $\mathbf{C}$ for the semiparametric model in the compact WLS-SVC formulation.

As previously stated, the same base of centroids $\mathbf{C}$ is employed in all the binary classifiers of the multiclass SVM.

It will be shown in Section IV that such a simple selection method leads to compact hybrid WLS-SVC/HMM recognition systems that are now capable of performing a real-time decoding of the speech utterances on a medium-complexity connected-digit recognition task. Namely, it will be shown in Table III that the complexity of the compact support vector machines (in terms of the number of centroids in the base) is 266 to 497 times lower than the complexity of the conventional SVMs (in terms of the number of support vectors).

IV. Experiments and Results

This section starts with a description of the experimental setup. Then, we present some experimental results that show the benefits of the proposed compact hybrid WLS-SVC/HMM system with respect to the baseline systems.

A. Database

1) Description: The well-known SpeechDat Spanish database [76] is used to assess the performance of the proposed system. This large vocabulary (more than $24,000$ words) continuous speech recognition database comprises recordings from $4,000$ Spanish speakers recorded at 8 kHz over the PSTN using an E-1 interface, in a noiseless office environment. This database comprises $160,000$ utterances with isolated and connected digits, natural numbers, spellings, city and company names, common application words, phonetically rich sentences, etc. Most items are read and some of them are spontaneously spoken.

The database is partitioned into three main sets: training set (80%), development or validation set (8%), and test set (12%). The original database is then processed to eliminate the *silence* samples placed at the beginning and end of the sentences, using the time marks in the database label files. As a result, the *training* set used for the baseline HMM-based

systems contains approximately 50 hours of continuous speech from 3,146 speakers (71,046 utterances).

The *development* set contains 7,436 utterances from 350 different speakers (5 hours of voice after preprocessing) with the same varied content as the training data set. A subset is used to select the word insertion log-probability for the Viterbi decoder, since we have found this value very sensitive to different noisy conditions, and the training parameters of the support vector machines (C, kernel parameters, growing and pruning thresholds, etc.).

The *test* set employed for validation corresponds to the connected-digit recognition task of the SpeechDat database, which comprises 2,122 utterances and 19,855 digits (5 hours of processed speech) from 499 different speakers. The number of recognized phonemes is restricted to the 18 present in Spanish digits (we have dropped the samples corresponding to the remaining 14 phonemes from the data set used for training the SVMs). The number of discarded samples represents just 8.8% of the samples in the whole training set.

Finally, it is worth noting that the experimental setup employed for validation is a trade-off between algorithmic approach suitable for continuous speech recognition and computational tractability using current SVM software implementation. First, the connected-digit task described in this work is set from a continuous speech recognition point of view, so that the proposed approach itself is scalable to more complex tasks. Second, the size of the SpeechDat database (50 hours of speech) allows us to investigate the different configurations described in the paper while extracting significant conclusions from the experimentation.

2) Database Contamination: The robustness of the hybrid SVM/HMM ASR systems has been tested in clean conditions and in the presence of additive noise. For that purpose, *white* and *babble* noises extracted from the NOISEX-92 database [77] were added to the clean speech signals at four different signal-to-noise ratios (SNRs), namely 12 dB, 9 dB, 6 dB, and 3 dB. Only the testing and development subsets have been corrupted in the way previously stated, whereas the acoustic models (GMMs and SVMs) have been estimated or trained using only clean speech.

3) Feature Extraction: We use a conventional parameterization based on 12 MFCCs (Mel-Frequency Cepstral Coefficients) plus the energy coefficient, and their first and second derivatives. Thus, a 39-dimensional feature vector is computed every 10 ms using an analysis window of 25 ms. We have employed the Cambridge University Hidden Markov Toolkit (HTK) [78] for this purpose.

The cepstral coefficients are then normalized on an utterance basis, a necessary task for noisy environments, where training and testing conditions do not match. Besides, this normalization is advisable to facilitate the convergence of SVMs. Thus, every parameter is normalized in mean and variance (CMVN) according to the following expression:

$$\hat{x}_t^{(i)} = \frac{x_t^{(i)} - \mu^{(i)}}{\sigma^{(i)}} \tag{25}$$

where $x_t^{(i)}$ represents the i^{th} component of the feature vector corresponding to frame t, and $\mu^{(i)}$ and $\sigma^{(i)}$ are the estimated mean and standard deviation from the whole utterance, respectively, for the i^{th} component.

4) Data Balancing and Context: As mentioned in Section III-C, the computational limitations of current SVM software implementations forced us to extract two reduced balanced data subsets (3 and 6 hours of speech) for the training of the SVMs. The new training datasets are extracted from the whole (non-balanced) SpeechDat training set by selecting phone samples randomly so that each class is equally represented. Consequently, these acoustic units appear in general contexts and not only in those observed in the test. In addition, a large percentage of the discarded samples correspond to silence segments (as silences represent approximately 34% of the original training set). Table I summarizes the distribution of data into these sets.

It is worth mentioning that hybrid speech recognition systems clearly benefit from the use of context information [29]. For scalability reasons, the use of context-dependent speech units is not straightforward. However, context information can be included in this case by joining adjacent feature vectors together into a single input vector, since SVMs can handle vectors of a very high dimensionality. The empirical study in the hybrid ANN/HMM ASR framework presented in [5] suggests an optimal context length of 3 frames, which roughly corresponds to the mean duration of the acoustic units (states of phoneme).

B. Baseline Systems

1) Baseline HMM Systems: A standard left-to-right HMM-based recognition system implemented using HTK, similar to that described in [79], is employed to produce a forced alignment necessary to obtain the labels for the SVMs, as SpeechDat is not phonetically labeled. More sophisticated techniques could be included in the recognizer, with minimal impact on the overall conclusions of this work.

Each of the 32 context-independent phone models consists of 3 active states (plus initial and final non-emitting states) where emission probabilities are modeled by a mixture of 32 Gaussians. The training process of the acoustic models consists of several steps, including an initial bootstrap models training, segmentation of the training set using those models, and iterative re-estimation of the parameters of the HMM.

This system is employed to produce the state-level segmentation of the training set used for the hybrid SVM/HMM systems, i.e., we label each frame with one of the possible 54 states (corresponding to 17 phones plus *silence*). To avoid the potential appearance of empty states, the HMM topology does not allow to obviate any of the states except in the */sil/* model whose central state is designed to model short pauses and allows a jump from the first emitting state to the last one and vice versa.

The results of a triphone-based HMM system are also included for the sake of completeness. This system, also based on [79], defines 5,357 triphone models with three active states modeling the emission probabilities by a mixture of 32 Gaussians.

The word error rates (WER) obtained for the baseline HMM recognizers in clean conditions are 2.41% for the

TABLE I
SUMMARY OF THE DATASETS EMPLOYED IN THE PAPER. THE THREE TRAINING SETS (NB–NON-BALANCED-, B1–BALANCED 1-, AND B2–BALANCED 2-) DIFFER IN THE PORTION OF THE AVAILABLE DATASET USED. DEVELOPMENT AND TEST SETS ARE THE SAME FOR ALL OF THE EXPERIMENTS.

Dataset	Train		Development		Test	
	# frames	Distribution	# frames	Distribution	# frames	Distribution
NB	16,378,624	Non-Balanced	1,682,065	Non-Balanced	1,656,102	Non-Balanced
B1	1,080,000	Balanced				
B2	2,160,000	Balanced				

monophone-based system and 1.87% for the triphone-based one. Previously published results on comparable tasks prove that the performances of our baseline HMM-based systems are in the state-of-the-art. Namely, the word error rate reported in [80] for a connected-digit recognition task using a triphone-based HMM system is 2.17%.

As we will show in a more detailed comparison in Section IV-C, the SVM-based ASR systems present similar or even better results in noisy conditions than the baseline monophone-based HMM system. However, there is still a gap with respect to the performance of context-dependent HMM systems. In our opinion, a more effective procedure for the SVM-based systems to take full benefit of contextual information should be developed in order to overcome this gap.

2) Baseline LibSVM/HMM System: A baseline hybrid SVM/HMM speech recognition system, based on the conventional formulation of the support vector machine, has been built for comparison purposes. This hybrid system is based on [11] and employs an SVM to estimate the HMM emission probabilities that will be used by a Viterbi decoder to obtain the transcription for the speech utterances.

The SVMs have been trained with the balanced data sets specified in Table I using the LibSVM toolbox. In this work we employ the versatile Gaussian kernel function:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right) \tag{26}$$

The optimal values for the training parameters C and σ were obtained empirically through a validation process that uses a subset of the development set described in Table I. Different values of the training parameters must be used for each of the context lengths, namely: $C = 1.0$, $\sigma = 32$ for a context length of 1 frame, and $C = 2.0$, $\sigma = 128$ for a context length of 3 frames.

The complexity of the SVM (samples that become support vectors with their corresponding Lagrange multipliers α) is determined by the training algorithm. In this case, the support vectors represent at least 69.87% and 65.36% of the balanced training sets when a context length of 1 and 3 frames is used, respectively. The dimension of the input feature vectors also depends on the context length employed in the experiments, leading to 39 and 117-dimensional training samples. The outputs of the SVM provide 54 *a posteriori* probabilities corresponding to each of the states in our system.

3) Compact WLS-SVC/HMM System: The structure of this system is similar to that of the previous one, but in this case the weighted least squares procedure described in Section III is used to train compact SVMs. A modified version of the software LibSVM has been employed for this purpose.

Different values for the training parameters were obtained for each context length through a validation process similar to that described before: $C = 8.0$, $\sigma = 128$, $\nu_g = 0.325$, $\nu_p = 0$ for a context length of 1 frame, and $C = 8.0$, $\sigma = 512$, $\nu_g = 0.21$, $\nu_p = 0$ for a context length of 3 frames.

As previously stated, the complexity of the compact SVM can be fixed *a priori* by imposing a semiparametric model on vector $\mathbf{w}$. The growing (ν_g) and pruning (ν_p) thresholds control the number of centroids that form the base for the model. Their optimal values result from a compromise between size and accuracy in the SVM. In this case, the centroids represent a maximum of 0.22% and 0.26% of the balanced training sets when a context length of 1 and 3 frames is used, respectively. It will be shown in the next section that such a huge reduction in the complexity of the support vector machine allows a real-time decoding of the speech utterances.

C. Experimental Results

Once the experimental setup has been described in detail in the previous sections, we proceed to present a performance comparison of the proposed compact WLS-SVC/HMM recognizer with both baseline LibSVM/HMM and standard HMM-based systems. Table II shows the word error rates obtained by these systems in a connected-digit recognition task. The conventional HMM-based recognition systems were trained using the whole non-balanced dataset (NB). For computational reasons, the hybrid SVM/HMM systems were trained using the two balanced subsets (B1 and B2) described in Table I. Context windows of 1 and 3 frames have been considered since previous works in the field have demonstrated the benefits derived from the inclusion of acoustic context in the hybrid approach.

In our opinion, the results in Table II show the potential of hybrid SVM/HMM systems. This fact is especially evident in noisy conditions, where the best LibSVM/HMM recognition system outperforms the baseline monophone-based HMM system. Furthermore, the improvements are statistically significant in five out of eight cases. The proposed compact WLS-SVC/HMM recognizer provides (statistically significant) better results than the monophone-based HMM system for *white* noise at 3 and 6 dB and equivalent performance (i.e., within the confidence intervals) for the remaining cases. It is worth mentioning that the proposed hybrid SVM/HMM systems provide competitive performance on this task, while using much less training samples, namely a maximum of 13% of the samples in the whole non-balanced dataset (NB)

TABLE II
PERFORMANCE COMPARISON OF HMM, LIBSVM/HMM AND WLS-SVC/HMM RECOGNITION SYSTEMS IN NOISY CONDITIONS. WORD ERROR RATES (WER) WITH 95% CONFIDENCE INTERVALS (CI) ARE SHOWN FOR DIFFERENT NOISE ENVIRONMENTS, ACOUSTIC CONTEXTS (1 AND 3 FRAMES) AND TRAINING DATASETS (B1, B2 AND NB).

Recognizer	Context	Trn. Set	Noise type & SNR							
			Clean	White				Babble			
				12 dB	9 dB	6 dB	3 dB	12 dB	9 dB	6 dB	3 dB
HMM-monophones	**1**	**NB**	2.41 ± 0.21	5.34 ± 0.31	7.13 ± 0.36	10.31 ± 0.42	14.77 ± 0.49	4.18 ± 0.28	6.36 ± 0.34	10.79 ± 0.43	18.26 ± 0.54
HMM-triphones	**1**	**NB**	1.87 ± 0.19	3.10 ± 0.24	4.37 ± 0.28	6.69 ± 0.35	10.32 ± 0.42	3.02 ± 0.24	4.67 ± 0.29	8.10 ± 0.38	14.96 ± 0.50
LibSVM/HMM	**1**	**B1**	2.82 ± 0.23	5.81 ± 0.33	7.90 ± 0.38	10.90 ± 0.43	15.72 ± 0.51	4.75 ± 0.30	6.84 ± 0.35	11.29 ± 0.44	19.40 ± 0.55
		B2	2.57 ± 0.22	5.34 ± 0.31	7.34 ± 0.36	10.43 ± 0.43	15.29 ± 0.50	4.54 ± 0.29	6.58 ± 0.34	10.90 ± 0.43	19.04 ± 0.55
	3	**B1**	2.42 ± 0.21	4.70 ± 0.29	6.32 ± 0.34	9.24 ± 0.40	14.05 ± 0.48	4.25 ± 0.28	6.36 ± 0.34	10.46 ± 0.43	17.48 ± 0.53
		B2	2.14 ± 0.20	4.42 ± 0.29	6.12 ± 0.33	9.02 ± 0.40	13.71 ± 0.48	4.09 ± 0.28	6.02 ± 0.33	9.95 ± 0.42	17.04 ± 0.52
WLS-SVC/HMM	**1**	**B1**	2.94 ± 0.23	5.45 ± 0.32	7.12 ± 0.36	9.93 ± 0.42	14.27 ± 0.49	4.71 ± 0.29	6.97 ± 0.35	11.01 ± 0.44	18.58 ± 0.54
		B2	2.74 ± 0.23	5.26 ± 0.31	7.01 ± 0.36	9.88 ± 0.42	14.21 ± 0.49	4.62 ± 0.29	6.66 ± 0.35	10.76 ± 0.43	18.81 ± 0.54
	3	**B1**	2.64 ± 0.22	4.96 ± 0.30	6.61 ± 0.35	9.49 ± 0.41	13.79 ± 0.48	4.68 ± 0.29	6.64 ± 0.35	10.53 ± 0.43	17.60 ± 0.53
		B2	2.62 ± 0.22	4.80 ± 0.30	6.47 ± 0.34	9.20 ± 0.40	13.51 ± 0.48	4.48 ± 0.29	6.51 ± 0.34	10.46 ± 0.43	17.49 ± 0.53

employed for training the Gaussian mixture models in the conventional HMM recognizers. Thus, support vector machines seem to be a promising future alternative to conventional acoustic modeling techniques in automatic speech recognition. However, it is worth noting that, nowadays, the proposed SVM-based systems benefit from the inclusion of a three-frame context window less than HMM-based systems do from the use of triphone models, as shown in Table II. Therefore, we think that more elaborate methods for the inclusion of contextual information in the SVM-based hybrid architecture are required to overcome current context-dependent HMM-based recognizers. As previously stated, several interesting alternatives can be found in [29]–[32].

Comparing now the two hybrid SVM-based systems, it can be said that the WLS-SVC/HMM recognition system proposed in this paper provides similar performance to the baseline LibSVM/HMM recognizer. Comparable word error rates are achieved by both systems, with the LibSVM/HMM system outperforming our proposal only in clean conditions (B2 training set and 3 frames context length). However, the compact WLS-SVC/HMM recognizer entails a much lower computational burden that allows it to perform a real-time decoding of the speech utterances. Table III shows a comparison of the decoding stage complexity for these systems.

The complexity of the acoustic models for each of the four speech recognition systems described above is determined by different sets of parameters. The model size for the baseline monophone-based HMM system ($1,728$ Gaussians) result from 54 states, each one modeled by a mixture of 32 Gaussians. The model size of the triphone-based HMM system is $123,776$ Gaussians, since several models share certain states. In the case of the hybrid LibSVM/HMM and WLS-SVC/HMM systems, their complexities are given by the number of support vectors in (6) and the number of centroids in (19), respectively. However, both values can be expressed in terms of the number of Gaussians to be evaluated at the decoding stage, due to the fact that such a kernel function has been employed in the SVMs. Table III shows how support vectors represent a large proportion of the training datasets in the case of the LibSVM-based system. This is an inherent result in speech recognition, where large datasets with highly overlapped classes lead to huge SVMs. In contrast, the weighted least squares training procedure allows us to impose a preset compact model that controls the size of the WLS-SVC. Consequently, the complexity of the acoustic model in the WLS-SVC/HMM system is reduced by between two and three orders of magnitude with respect to that of the baseline hybrid recognizer.

Decoding times for the speech recognition systems, referenced to real-time (RT) performance, are also presented in Table III[1]. From these results, it can be seen that the proposed compact WLS-SVC/HMM recognizer achieves similar performance to the baseline LibSVM/HMM system with a much lower complexity. The reduction in decoding time is proportional to the reduction of the model sizes. Although these recognition times are still higher than those of the conventional HMM-based systems, the proposed hybrid recognizer is able to perform a real-time decoding of the test set in three out of

[1]Due to the huge computational burden of the LibSVM/HMM recognizer, all of the decoding time measures in Table III were taken over a reduced test set and then extrapolated. For this purpose, a PC equipped with an Intel Core 2 Duo E8400 processor at 3 GHz and 3 GB of RAM was employed. Nonetheless, word error rates shown in Table II were obtained over the whole test set. For the LibSVM/HMM case, a computer grid was employed.

TABLE III
COMPARISON OF THE COMPLEXITY OF HMM, LIBSVM/HMM AND WLS-SVC/HMM RECOGNITION SYSTEMS AT THE DECODING STAGE. ACOUSTIC MODEL SIZES, IN TERMS OF THE NUMBER OF GAUSSIAN FUNCTIONS TO EVALUATE, AND DECODING TIMES, REFERENCED TO REAL-TIME (RT) PERFORMANCE, ARE PRESENTED FOR ALL OF THE RECOGNITION SYSTEMS.

Recognizer	Context	Trn. Set	Size (# Gaussians)	Decoding Time (xRT)
HMM-monophones	**1**	**NB**	1,728	0.08
HMM-triphones	**1**	**NB**	123,776	0.13
LibSVM/HMM	**1**	**B1**	790,138	25.03
		B2	1,509,230	47.36
	3	**B1**	748,671	50.64
		B2	1,411,881	not available
WLS-SVC/HMM	**1**	**B1**	2,346	0.75
		B2	3,039	0.81
	3	**B1**	2,814	0.97
		B2	3,674	1.31

four cases. It is worth noting that although the complexities of the HMM-based systems are similar or even higher than those of the WLS-SVC/HMM systems, the decoding time of the latter is considerably higher. The reason is that all of the Gaussian kernels must be evaluated in the multiclass SVM to obtain a single posterior probability. On the other hand, only those models corresponding to active nodes in the Viterbi search must be evaluated at a given time in the HMM-based systems.

Finally, we would like to highlight a subset of results selected from Table II. For that purpose these results are replicated graphically in Fig. 1. First, let us compare the results achieved by the WLS-SCV/HMM system for the two training databases (B1 and B2). Although the size of the training database has a notable influence on the decoding complexity, since it determines the number of centroids, the differences in performances are small and not statistically significant. Therefore, the proposed system can be trained using a really small database. Second, since the contextual information has a noticeable influence on the system performance, we focus our attention on the results achieved using a three-frame context window (denoted as w3). We can see that the proposed WLS-SVC/HMM system attains competitive performance with respect to the monophone-based HMM system in both clean and noisy conditions, while reducing the complexity of the SVM/HMM system enough to allow for real-time speech recognition. In our opinion, these results represent an important step forward for SVM-based speech recognition, although further research in this framework is still required to allow for practical application of the proposed system in more demanding ASR tasks.

V. CONCLUSION AND FURTHER WORK

The hybrid speech recognition framework has demonstrated its capability to overcome some of the limitations of HMM-based recognizers. Support vector machines have several advantages over classical artificial neural networks, especially in noisy conditions. However, their computational burden has prevented them from being used in practice in ASR, although several preliminary hybrid systems can be found in the literature [10]–[12]. In this paper, we suggest the use of a weighted least squares training procedure [14] that allows us to control the complexity of the resulting SVM (denoted as WLS-SVC) by imposing a preset compact model. Other practical issues related to the application of SVMs in automatic speech recognition are also addressed. An exhaustive experimental study based on a connected-digit recognition task reveals the proposed hybrid WLS-SVC/HMM recognizer as a promising starting point for the development of preliminary SVM-based ASR systems. Specifically, we would like to highlight the following conclusions:

- Competitive performance with respect to standard monophone-based HMM systems has been obtained in clean and noisy conditions. Furthermore, statistically significant better results have been obtained in a few noisy cases.
- Real-time speech decoding has been achieved by means of compact support vector machines.
- Only a small subset (from 6.5% to 13%) of the full training set (NB) is required to obtain competitive results on the selected task, which partially alleviates the inherent complexity of SVMs at the training stage.

Once we have implemented a first hybrid WLS-SVC/HMM system that is able to perform real-time speech decoding on a medium-complexity connected-digit recognition task, further research lines open up in order to improve its performance and to extend it to large-scale ASR. In particular:

- Development of better procedures for the selection of the base of centroids for the compact WLS-SVC in order to obtain larger reductions in the complexity of the SVM-based recognizers and better recognition performance.
- Analysis of more suitable multiclass architectures and probability estimation methods for the speech recognition problem at hand.
- Use of more adequate spectral feature representations and adoption of more elaborate methods for the inclusion of contextual information in the hybrid architecture such as [29]–[32], which should contribute to overcome the gap with respect to triphone-based HMM systems.

Finally, the use of specific techniques for sequence data prediction such as *structured SVMs* (e.g. Hidden Markov Support Vector Machines [81], Maximum Margin Markov

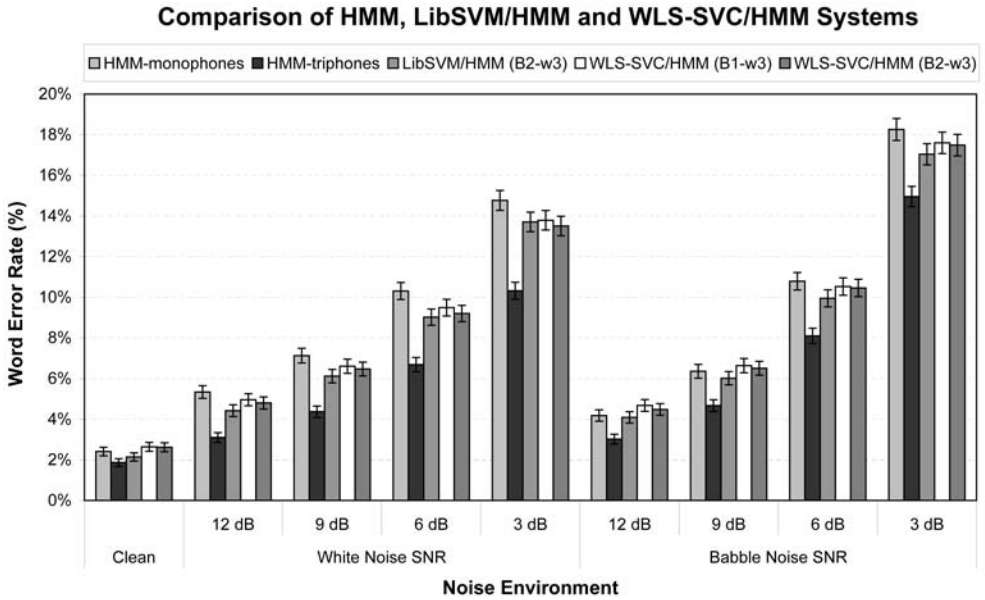

Fig. 1. Comparison of the most remarkable results for HMM, LibSVM/HMM and WLS-SVC/HMM recognition systems in noisy conditions. Abbreviation w3 denotes three-frame context length. B1 and B2 denote the training datasets. Vertical segments represent 95% confidence intervals (CI).

Networks [82], Kernel Conditional Graphical Models [83]) is also among our future research lines.

ACKNOWLEDGMENT

The authors would like to thank Prof. Dr. Ángel Navia-Vázquez for his advice on the compact WLS-SVC formulation and the anonymous reviewers for their comments and suggestions.

REFERENCES

[1] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: a Hybrid Approach*. Norwell, MA, USA: Kluwer Academic Publishers, 1993.

[2] N. Morgan and H. Bourlard, "Continuous Speech Recognition: an introduction to the hybrid HMM/connectionist approach," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 24–42, May 1995.

[3] E. Trentin and M. Gori, "A Survey of Hybrid ANN/HMM Models for Automatic Speech Recognition," *Neurocomputing*, vol. 37, no. 1-4, pp. 91–126, Apr. 2001.

[4] P. Pujol, S. Pol, C. Nadeu, A. Hagen, and H. Bourlard, "Comparison and Combination of Features in a Hybrid HMM/MLP and a HMM/GMM Speech Recognition System," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 14–22, Jan. 2005.

[5] A. I. García-Moral, R. Solera-Ureña, C. Peláez-Moreno, and F. Díaz-de-María, "Data Balancing for Efficient Training of Hybrid ANN/HMM Automatic Speech Recognition Systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 468–481, Mar. 2011.

[6] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[7] H. Shimodaira, K. Noma, M. Nakai, and S. Sagayama, "Support Vector Machine with Dynamic Time-Alignment Kernel for Speech Recognition," in *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, Aalborg, Denmark, Sep. 2001, pp. 1841–1844.

[8] N. D. Smith and M. Niranjan, "Data-dependent kernels in SVM classification of speech patterns," in *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, vol. 1, Beijing, China, Oct. 2000, pp. 297–300.

[9] A. Ganapathiraju, J. E. Hamaker, and J. Picone, "Hybrid SVM/HMM architectures for speech recognition," in *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000)*, vol. 4, Beijing, China, Oct. 2000, pp. 504–507.

[10] S. E. Krüger, M. Schafföner, M. Katz, E. Andelic, and A. Wendemuth, "Speech Recognition with Support Vector Machines in a Hybrid System," in *Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech 2005)*, Lisbon, Portugal, Sep. 2005, pp. 993–996.

[11] J. Padrell-Sendra, D. Martín-Iglesias, and F. Díaz-de-María, "Support vector machines for continuous speech recognition," in *Proceedings of the 14th European Signal Processing Conference (EUSIPCO 2006)*, Florence, Italy, Sep. 2006.

[12] D. Bolaños-Alonso, "Advances in the Application of Support Vector Machines as Probabilistic Estimators for Continuous Automatic Speech Recognition," Ph.D. Thesis, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Madrid, Spain, Nov. 2008. [Online]. Available: http://digitool-uam.greendata.es//exlibris/dtl/d3_1/apache_media/16328.pdf

[13] M. J. F. Gales and F. Flego, "Discriminative classifiers with adaptive kernels for noise robust speech recognition," *Computer Speech & Language*, vol. 24, no. 4, pp. 648–662, Oct. 2010.

[14] F. Pérez-Cruz, A. Navia-Vázquez, J. L. Rojo-Álvarez, and A. Artés-Rodríguez, "A new training algorithm for support vector machines," in *Proceedings of the Fifth Bayona Workshop on Emerging Technologies in Telecommunications*, Baiona, Spain, 1999, pp. 116–120. [Online]. Available: http://www.tsc.uc3m.es/~fernando/research3.html

[15] F. Pérez-Cruz, "Máquina de Vectores Soporte Adaptativa y Compacta," Ph.D. Thesis, Universidad Politécnica de Madrid, Madrid, Spain, Oct. 2000. [Online]. Available: http://www.tsc.uc3m.es/~fernando/tesis.ps.zip

[16] A. Navia-Vázquez, F. Pérez-Cruz, A. Artés-Rodríguez, and A. R. Figueiras-Vidal, "Weighted Least Squares Training of Support Vector Classifiers Leading to Compact and Adaptive Schemes," *IEEE Transactions on Neural Networks*, vol. 12, no. 5, pp. 1047–1059, Sep. 2001.

[17] Y. Engel, S. Mannor, and R. Meir, *Machine Learning: ECML 2002*, ser. Lecture Notes in Computer Science (LNCS). Berlin/Heidelberg, Germany: Springer-Verlag, 2002, vol. 2430/2002, ch. Sparse Online Greedy Support Vector Regression, pp. 84–96.

[18] ——, "The Kernel Recursive Least-Squares Algorithm," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2275–2285, Aug. 2004.

[19] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Upper Saddle River, NJ, USA: Prentice Hall, 1998.

[20] H. Bourlard and N. Morgan, *Adaptive Processing of Sequences and Data Structures*, ser. Lecture Notes in Computer Science (LNCS). Berlin/Heidelberg, Germany: Springer-Verlag, 1998, vol. 1387/1998, ch.

Hybrid HMM/ANN Systems for Speech Recognition: Overview and New Research Directions, pp. 389–417.

[21] J. P. Neto, C. Martins, and L. B. Almeida, "An incremental speaker-adaptation technique for hybrid HMM-MLP recognizer," in *Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP 1996)*, vol. 3, Philadelphia, PA, USA, Oct. 1996, pp. 1293–1296.

[22] ——, "Speaker-adaptation in a hybrid HMM-MLP recognizer," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1996)*, vol. 6, Washington, DC, USA, May 1996, pp. 3382–3385.

[23] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. De Mori, "Linear hidden transformations for adaptation of hybrid ANN/HMM models," *Speech Communication*, vol. 49, no. 10-11, pp. 827–835, Oct. 2007.

[24] S. Thomas, S. Ganapathy, and H. Hermansky, "Spectro-Temporal Features for Automatic Speech Recognition using Linear Prediction in Spectral Domain," in *Proceedings of the 16th European Signal Processing Conference (EUSIPCO 2008)*, Lausanne, Switzerland, Aug. 2008.

[25] S. Ganapathy, S. Thomas, and H. Hermansky, "Modulation frequency features for phoneme recognition in noisy speech," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. EL8–EL12, Jan. 2009.

[26] J. Frankel and S. King, "A Hybrid ANN/DBN Approach to Articulatory Feature Recognition," in *Proceedings of the 9th International Conference on Speech Communication and Technology (Interspeech 2005)*, Lisbon, Portugal, Sep. 2005, pp. 3045–3048.

[27] J. Frankel, M. Wester, and S. King, "Articulatory feature recognition using dynamic Bayesian networks," *Computer Speech & Language*, vol. 21, no. 4, pp. 620–640, Oct. 2007.

[28] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, Feb. 2007.

[29] N. Morgan, B. Y. Chen, Q. Zhu, and A. Stolcke, "Scaling up: Learning large-scale recognition methods from small-scale recognition tasks," in *Proceedings of the Special Workshop in Maui (SWIM)*, Hawaii, USA, Jan. 2004.

[30] S. Y. Zhao, S. Ravuri, and N. Morgan, "Multi-Stream to Many-Stream: Using Spectro-Temporal Features for ASR," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, Brighton, UK, Sep. 2009, pp. 2951–2954.

[31] A. Abad and J. Neto, "Incorporating acoustical modelling of phone transitions in an hybrid ANN/HMM speech recognizer," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech 2008)*, Brisbane, Australia, Sep. 2008, pp. 2394–2397.

[32] A. Abad, T. Pellegrini, I. Trancoso, and J. Neto, "Context dependent modelling approaches for hybrid speech recognizers," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, Chiba, Japan, Sep. 2010, pp. 2950–2953.

[33] R. Solera-Ureña, D. Martín-Iglesias, A. Gallardo-Antolín, C. Peláez-Moreno, and F. Díaz-de-María, "Robust ASR using Support Vector Machines," *Speech Communication*, vol. 49, no. 4, pp. 253–267, Apr. 2007.

[34] R. Solera-Ureña, J. Padrell-Sendra, D. Martín-Iglesias, A. Gallardo-Antolín, C. Peláez-Moreno, and F. Díaz-de-María, *Progress in Nonlinear Speech Processing*, ser. Lecture Notes in Computer Science (LNCS). Berlin/Heidelberg, Germany: Springer-Verlag, 2007, vol. 4391/2007, ch. SVMs for Automatic Speech Recognition: A Survey, pp. 190–216.

[35] A. Ech-Cherif, M. Kohili, A. Benyettou, and M. Benyettou, "Lagrangian support vector machines for phoneme classification," in *Proceedings of the 9th International Conference on Neural Information Processing (ICONIP 2002)*, vol. 5, Singapore, Nov. 2002, pp. 2507–2511.

[36] S. V. Gangashetty, C. Chandra-Sekhar, and B. Yegnanarayana, "Combining Evidence from Multiple Classifiers for Recognition of Consonant-Vowel Units of Speech in Multiple Languages," in *Proceedings of 2005 International Conference on Intelligent Sensing and Information Processing*, Chennai, India, Dec. 2005, pp. 387–391.

[37] D. Martín-Iglesias, J. Bernal-Chaves, C. Peláez-Moreno, A. Gallardo-Antolín, and F. Díaz-de-María, *Nonlinear analyses and algorithms for speech processing*, ser. Lecture Notes in Computer Science (LNCS). Berlin/Heidelberg, Germany: Springer-Verlag, 2005, vol. 3817/2005, ch. A Speech Recognizer Based on Multiclass SVMs with HMM-guided Segmentation, pp. 256–266.

[38] P. Clarkson and P. J. Moreno, "On the use of support vector machines for phonetic classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1999)*, vol. 2, Phoenix, AZ, USA, Mar. 1999, pp. 585–588.

[39] N. Thubthong and B. Kijsirikul, "Support vector machines for Thai phoneme recognition," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 9, no. 6, pp. 803–813, 2001.

[40] C. Chandra-Sekhar, W. F. Lee, K. Takeda, and F. Itakura, "Acoustic modeling of subword units using support vector machines," in *Proceedings of the Workshop on Spoken Language Processing (WSLP 2003)*, Mumbai, India, Jan. 2003, pp. 79–86.

[41] A. Ganapathiraju, "Support vector machines for speech recognition," Ph.D. Thesis, Mississipi State University, Mississippi, MS, USA, Jan. 2002. [Online]. Available: http://www.isip.piconepress.com/publications/books/msstate_theses/2002/support_vectors/thesis/thesis_final.pdf

[42] A. Ganapathiraju, J. E. Hamaker, and J. Picone, "Applications of Support Vector Machines to Speech Recognition," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2348–2355, Aug. 2004.

[43] S. Fine, G. Saon, and R. A. Gopinath, "Digit recognition in noisy environments via a sequential GMM/SVM system," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002)*, vol. 1, Orlando, FL, USA, May 2002, pp. 49–52.

[44] M. Schafföner, S. E. Krüger, E. Andelic, M. Katz, and A. Wendemuth, "Limited Training Data Robust Speech Recognition Using Kernel-Based Acoustic Models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, vol. 1, Toulouse, France, May 2006, pp. 1137–1140.

[45] D. Bolaños and W. Ward, "Implicit State-Tying for Support Vector Machines Based Speech Recognition," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (Interspeech 2008)*, Brisbane, Australia, Sep. 2008, pp. 924–927.

[46] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," in *Proceedings of the fifth Annual Workshop on Computational Learning Theory*, Pittsburgh, PA, USA, Jul. 1992, pp. 144–152.

[47] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley-Interscience, 1998.

[48] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York, NY, USA: Springer, 1999.

[49] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer, "Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning," *Automation and Remote Control*, vol. 25, pp. 821–837, 1964.

[50] J. Weston and C. Watkins, "Multi-Class Support Vector Machines," Department of Computer Science, Royal Holloway, University of London, Egham, United Kingdom, Tech. Rep., 1998.

[51] K. Crammer and Y. Singer, "On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines," *Journal of Machine Learning Research*, vol. 2, no. 5, pp. 265–292, 2001.

[52] Y. Lee, Y. Lin, and G. Wahba, "Multicategory Support Vector Machines: Theory and Application to the Classification of Microarray Data and Satellite Radiance Data," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 67–81, Mar. 2004.

[53] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large Margin DAGs for Multiclass Classification," in *Advances in Neural Information Processing Systems 12*, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds. Cambridge, MA, USA: MIT Press, 2000, pp. 547–553.

[54] T. G. Dietterich and G. Bakiri, "Solving Multiclass Learning Problems via Error-Correcting Output Codes," *Journal of Artificial Intelligence Research*, vol. 2, no. 1, pp. 263–286, 1995.

[55] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing Multiclass to Binary: a Unifying Approach for Margin Classifiers," *Journal of Machine Learning Research*, vol. 1, pp. 113–141, 2000.

[56] C. W. Hsu and C.-J. Lin, "A Comparison of Methods for Multiclass Support Vector Machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, Mar. 2002.

[57] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[58] J. C. Platt, *Advances in Large Margin Classifiers*. Cambridge, MA, USA: MIT Press, 2000, ch. Probabilities for SV Machines, pp. 61–74.

[59] T.-F. Wu, C.-J. Lin, and R. C. Weng, "Probability Estimates for Multi-class Classification by Pairwise Coupling," *Journal of Machine Learning Research*, vol. 5, pp. 975–1005, 2004.

[60] V. García, J. Sánchez, and R. Mollineda, *Progress in Pattern Recognition, Image Analysis and Applications*, ser. Lecture Notes in Computer Science (LNCS). Berlin/Heidelberg, Germany: Springer-Verlag, 2008,

vol. 4756/2008, ch. An empirical study of the behavior of classifiers on imbalanced and overlapped data sets, pp. 397–406.
[61] G. M. Weiss, B. Zadrozny, and M. Saar-Tsechansky, "Guest editorial: special issue on utility-based data mining," *Data Mining and Knowledge Discovery*, vol. 17, no. 2, pp. 129–135, Oct. 2008.
[62] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory Under-sampling for Class-Imbalance Learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 2, pp. 539–550, Apr. 2009.
[63] L. Tóth and A. Kocsor, *Artificial Neural Networks: Biological Inspirations - ICANN 2005*, ser. Lecture Notes in Computer Science (LNCS). Berlin/Heidelberg, Germany: Springer-Verlag, 2005, vol. 3696/2005, ch. Training HMM/ANN Hybrid Speech Recognizers by Probabilistic Sampling, pp. 597–603.
[64] S. Scanzio, P. Laface, R. Gemello, and F. Mana, "Speeding-Up Neural Network Training Using Sentence and Frame Selection," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, Antwerp, Belgium, Aug. 2007, pp. 1725–1728.
[65] D. Albesano, R. Gemello, and F. Mana, "Hybrid HMM-NN for speech recognition and prior class probabilities," in *Proceedings of the 9th International Conference on Neural Information Processing (ICONIP 2002)*, vol. 5, Singapore, Nov. 2002, pp. 2391–2395.
[66] A. Hagen, "Robust speech recognition based on multi-stream processing," Ph.D. Thesis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, Dec. 2001. [Online]. Available: http://infoscience.epfl.ch/search.py?recid=32973
[67] F. Pérez-Cruz, A. Navia-Vázquez, P. L. Alarcón-Diana, and A. Artés-Rodríguez, "SVC-based equalizer for burst TDMA transmissions," *Signal Processing*, vol. 81, no. 8, pp. 1681–1693, Aug. 2001.
[68] F. Pérez-Cruz, C. Bousoño-Calzón, and A. Artés-Rodríguez, "Convergence of the IRWLS Procedure to the Support Vector Machine Solution," *Neural Computation*, vol. 17, no. 1, pp. 7–18, Jan. 2005.
[69] H. W. Kuhn and A. W. Tucker, "Nonlinear programming," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, USA, Jul.-Aug. 1950, pp. 481–492.
[70] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.
[71] A. N. Tikhonov and V. Y. Arsenin, *Solution of ill-posed problems*. Washington, DC, USA: W. H. Winston & Sons, 1977.
[72] G. Kimeldorf and G. Wahba, "Some Results on Tchebycheffian Spline Function," *Journal of Mathematical Analysis and Applications*, vol. 33, no. 1, pp. 82–95, Jan. 1971.
[73] C. J. C. Burges and B. Schölkopf, "Improving the Accuracy and Speed of Support Vector Machines," in *Advances in Neural Information Processing Systems 9*, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. Cambridge, MA, USA: MIT Press, 1997, pp. 375–381.
[74] S. E. Borys, "An SVM Front End Landmark Speech Recognition System," M.Sc. Thesis, Graduate College, University of Illinois at Urbana-Champaign, Urbana, IL, USA, 2008. [Online]. Available: http://www.isle.illinois.edu/~sborys/BorysMSThesis08.pdf
[75] F. Orabona, J. Keshet, and B. Caputo, "Bounded Kernel-Based Online Learning," *Journal of Machine Learning Research*, vol. 10, pp. 2643–2666, 2009.
[76] A. Moreno, "SpeechDat Spanish Database for Fixed Telephone Network," Universitat Politecnica de Catalunya, Tech. Rep., 1997.
[77] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.
[78] S. Young, G. Evermann, M. Gales, T. Hain, and D. Kershaw, *HTK-Hidden Markov Model toolkit (ver. 3.4)*, Cambridge University, 2006.
[79] F. T. Johansen, N. Warakagoda, B. Lindberg, G. Lehtinen, Z. Kačič, A. Žgank, K. Elenius, and G. Salvi, "The COST 249 SpeechDat Multilingual Reference Recogniser," in *COST 249 MCM, Technical Annex*, 1999.
[80] B. Kotnik, Z. Kačič, and B. Horvat, "Development and Integration of the LDA-Toolkit into the COST249 SpeechDat (II) SIG Reference Recognizer," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, May 2004.
[81] Y. Altun, I. Tsochantaridis, and T. Hofmann, "Hidden Markov Support Vector Machines," in *Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003)*, Washington, DC, USA, Aug. 2003, pp. 3–10.
[82] B. Taskar, C. Guestrin, and D. Koller, "Max-Margin Markov Networks," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA, USA: MIT Press, 2004.
[83] F. Pérez-Cruz, Z. Ghahramani, and M. Pontil, "Kernel conditional graphical models," in *Predicting Structured Data*, F. Bakir, T. Hofmann, B. Schölkopf, A. J. Smola, B. Taskar, and S. V. N. Vishwanathan, Eds. Cambridge, MA, USA: MIT Press, 2007, pp. 265–282.

Rubén Solera-Ureña received the Telecommunication Engineering degree in 2004 and the Ph.D. degree in 2011 from the University Carlos III of Madrid, Spain. Currently, he is an Assistant Professor at the Department of Signal Theory and Communications, University Carlos III of Madrid, Spain.

His primary research interests include multimedia processing, robust speech recognition, nonlinear speech processing and machine learning for signal processing, with special emphasis in speech applications.

Ana Isabel García-Moral received her Telecommunication Engineering degree from the University Carlos III of Madrid in 2003. From 2007 to 2009, she has been an Assistant Professor at the Department of Signal Theory and Communications, University Carlos III of Madrid. From 2009 she has been working in different private companies related to speech recognition and semantic classification based in Nuance technology and is currently pursuing a Ph.D. from the aforementioned university.

Her research interests include natural language processing, speech recognition, machine learning, pattern recognition and classification.

Carmen Peláez-Moreno received her Telecommunication Engineering degree from the Public University of Navarre in 1997 and Ph.D. from the University Carlos III of Madrid in 2002. Her Ph.D. thesis has been awarded a 2002 Best Doctoral Thesis Prize from the Spanish Official Telecommunication Engineering Association (COIT-AEIT).

From March to Dec. 2004, she participated in the International Computer Science Institute's (ICSI, Berkeley (CA)) Fellowship Program. Since Nov. 2009, she is an Associate Professor in the Department of Signal Theory and Communications at the University Carlos III of Madrid. Her research interests include robust speech recognition and perception, video and speech coding, multimedia information retrieval, machine learning and data analysis. She has co-authored several papers in prestigious international journals, books and peer-reviewed conferences.

Manel Martínez-Ramón received his degree in Telecommunication Engineering in 1994 (Universitat Politècnica de Catalunya, Spain), and finished his Ph.D., also in Telecommunication Engineering in 1999 (University Carlos III of Madrid, Spain). He is with the Department of Signal Theory and Communications, University Carlos III of Madrid.

His research topics are applications of the statistical learning to signal processing, with emphasis in communications, speech, and brain imaging. He has coauthored about 25 papers in international journals and 40 conference papers on these topics. He has written a book on applications of SVMs to antennas and electromagnetics and coauthored several book chapters. He is a senior member of the IEEE since 2004.

Fernando Díaz-de-María received the Telecommunication Engineering degree in 1991 and the Ph.D. degree in 1996 from the Polytechnic University of Madrid, Spain. From Oct. 1996, he is an Associate Professor at the Department of Signal Theory and Communications, University Carlos III of Madrid, Spain. From Oct. 97, he has held several offices in both, his Department and his University.

His primary research interests include robust speech processing, image and video analysis, and image and video coding. He has led numerous Projects and Contracts in the mentioned fields. He is co-author of several papers in prestigious international journals, chapters in international books and quite a few papers in revised national and international conferences.